1.0 PURPOSE AND BACKGROUND OF PERFORMANCE STANDARDS

1.1 Introduction

Prior to the acceptance of a new test method for regulatory testing applications, validation studies are conducted to assess its reliability (i.e., the extent of intra- and inter-laboratory reproducibility) and its accuracy (i.e., the ability of the test method to correctly predict or measure the biological effect of interest) (ICCVAM 1997, 1999, 2002; OECD 1996, 2002a). The purpose of performance standards is to communicate the basis by which new proprietary (i.e., copyrighted, trademarked, registered) and nonproprietary test methods have been determined to have sufficient accuracy and reliability for specific testing purposes. These performance standards, based on test methods accepted by regulatory agencies, can be used to evaluate the reliability and accuracy of other test methods that are based on similar scientific principles and measure or predict the same biological or toxic effect. EpiDerm™, a human skin model system for skin corrosion, is an example of a test method that underwent an expedited validation process because it was mechanistically and functionally similar to a previously validated human skin model system for skin corrosion called EPISKIN™ (see Section 3.0). Another example of their application would be to evaluate the use of mouse or human skin rather than rat skin in the rat skin transcutaneous electrical resistance (TER) assay. This section describes the three elements of performance standards identified by ICCVAM (ICCVAM 2003), the ICCVAM process for developing performance standards during test method evaluations, and the ICCVAM process to retrospectively develop performance standards for previously reviewed test methods. A retrospecitive process was used to develop performance standards for three types of validated *in vitro* corrosivity test methods (a noncellular membrane barrier test system, cultured human skin model systems, the rat skin TER test method), and those perfomance standards are provided in **Sections 2.0**, **3.0**, and **4.0**.

1.2 Elements of ICCVAM Performance Standards

The three elements of performance standards are:

- **Essential test method components:** These consist of essential structural, functional, and procedural elements of a validated test method that should be included in the protocol of a proposed mechanistically and functionally similar test method. Essential test method components include unique characteristics of the test method, critical procedural details, and quality control measures. Adherence to essential test method components will help to assure that a proposed test method is structurally and functionally similar to the corresponding validated test method.

- **A minimum list of reference chemicals:** Reference chemicals are used to assess the accuracy and reliability of a proposed mechanistically and functionally similar test method. These chemicals are a representative subset of those used to demonstrate the reliability and the accuracy of the validated test method. To the extent possible, this subset of chemicals should:
  - be representative of the range of responses that the validated test method is capable of measuring or predicting
  - have produced consistent results in the validated test method and the *in vivo* reference test method and/or the species of interest

- have well-defined chemical structures
- be readily available
- not be associated with excessive hazard or prohibitive disposal costs
- have performance characteristics (e..g., accuracy, sensitivity, specificity, false negative and false positive rates) in the validated test method that approximate the performance values obtained for all appropriate substances during the validation process

These reference chemicals are the minimum number that should be used to evaluate the performance of a proposed mechanistically and functionally similar test method. Reference chemicals should not be used to develop the prediction model for the proposed test method. If any of the recommended reference chemicals are not available, other chemicals for which adequate reference data are available could be substituted. To the extent possible, the substituted chemical(s) should be of the same chemical class as the original reference chemical(s). If desired, additional chemicals representing other chemical or product classes and for which adequate reference data are available can be used to more comprehensively evaluate the accuracy of the proposed test method. However, none of these additional chemicals should have been used to develop the proposed test method.

- **Accuracy and reliability values:** These are the accuracy and reliable characteristics that the proposed test method should be comparable to when evaluated using the minimum list of reference chemicals.

1.3 ICCVAM Process for the Development of Performance Standards

The process followed by ICCVAM for developing performance standards for new test methods is as follows (ICCVAM, 2003):

- NICEATM and the appropriate ICCVAM working group develop proposed performance standards for consideration during the ICCVAM evaluation process. If performance standards are proposed by a test method sponsor, they will be considered by ICCVAM at this stage. Generally, the proposed performance standards are based on the information and data provided in the test method submission or other available applicable data.
- The ICCVAM/NICEATM Peer Review Panel evaluates the proposed performance standards for completeness and appropriateness during its evaluation of the validation status of the proposed test method. The proposed performance standards, as well as the test method submission, are made available to the public for comment prior to and during the Peer Review Panel meeting.
- The appropriate ICCVAM working group, with the assistance of NICEATM, prepares the final performance standards for ICCVAM approval, taking into consideration the recommendations of the Peer Review Panel and public comments.

Performance standards recommended by ICCVAM are incorporated into ICCVAM test method evaluation reports, which are then provided to U.S. Federal agencies and made available to the public. Regulatory authorities can then reference the performance standards in the ICCVAM report when they communicate their acceptance of a new test method. In addition, performance standards adopted by U.S. Federal regulatory authorities can be provided in guidelines issued for new test

methods. Availability of ICCVAM test method evaluation reports are announced routinely in the *Federal Register*, NTP Newsletters, and by e-mail to ICCVAM/NICEATM listserve groups.

1.4 ICCVAM Development of Recommended Performance Standards for *In Vitro* Test Methods for Skin Corrosion

Skin corrosion refers to the visible destruction or irreversible alteration of skin following exposure of the skin to a chemical substance. Skin corrosivity has traditionally been assessed by applying the test substance to the skin of living animals and evaluating the extent of tissue damage after a fixed period of time (OECD 2002b; EPA 1998). Some U.S. regulatory authorities require determination of corrosivity using three categories of responses, as provided in **Table 1-1** (EPA 1998; DOT 2003a, 2003b).

Table 1-1 Skin Corrosive Category and Subcategories

Corrosive Category (category 1) (applies to authorities not using subcategories)	Potential Corrosive Subclasses[1] (UN Packing Group Classification[2])	Corrosive in at least 1 of 3 animals	
		Exposure	Observation
Corrosive	Corrosive subcategory 1A (I)	≤3 minutes	≤1 hour
	Corrosive subcategory 1B (II)	>3 minutes / ≤1hour	≤14 days
	Corrosive subcategory 1C (III)	>1 hour / ≤4 hours	≤14 days

[1] Classifications designated by the United Nations (UN) Globally Harmonised System for the Classification and Labelling of Chemical Substances and Mixtures (GHS) (UN 2003a).

[2] Corresponding UN packing group classifications to be used for the transport of dangerous goods (UN 2003b).

The EPA test guideline (EPA 1996), a DOT exemption (DOT 2002), and a globally-harmonized tiered testing strategy (UN 2003a) for the assessment of skin corrosivity allow for the use of validated and accepted *in vitro* methods. In both the EPA guideline and the tiered testing strategy, positive results from *in vitro* test methods can be used to classify a substance as corrosive without the need for animal testing. Substances that are negative *in vitro* should undergo additional testing in accordance with the tiered testing strategy. The DOT exemption allows for the determination of corrosivity and noncorrosivity of specific classes of chemicals for certain transport testing circumstances. The use of *in vitro* methods to identify corrosive substances can therefore avoid pain and distress that may result from the application of corrosive substances to animals.

A number of *in vitro* test methods have been proposed as alternatives for the standard *in vivo* rabbit skin procedure to identify corrosive substances. Generally, these test methods have involved the use of a noncellular membrane barrier test system, cultured human skin model systems, or isolated rat skin (Fentem et al. 1998). ICCVAM previously evaluated and recommended four validated test methods for assessing the dermal corrosivity hazard potential of substances:

Corrositex®, EPISKIN™, EpiDerm™ (EPI-200), and the rat skin TER Assay (ICCVAM 1999, 2002). Subsequently, the EPA requested that ICCVAM establish performance standards for the three proprietary *in vitro* dermal corrosivity test methods (Corrositex®, EPISKIN™, EpiDerm™ [EPI-200]) and the non-proprietary rat skin TER test method. In response, the ICCVAM Dermal Corrosivity and Irritation Working Group (DCIWG) drafted proposed performance standards based on these validated *in vitro* test methods. As described earlier in this section, for future test methods evaluated by ICCVAM, performance standards will be included as part of the test method recommendations forwarded to U.S. Federal regulatory authorities.

In a *Federal Register* notice published on July 1, 2003**,** NICEATM announced the availability of and invited public comment on the proposed performance standards for the three types of validated *in vitro* test methods for assessing the dermal corrosivity hazard potential of chemicals (**Appendix A**). Public comments were received from individuals representing five organizations (**Appendix B**). Comments on the draft document were also obtained during public meetings of the ICCVAM Scientific Advisory Committee on Alternative Toxicological Methods (SACATM) in August, 2003, and the EPA Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) Scientific Advisory Panel (SAP) in October, 2003 (Appendix C). These comments, which are discussed briefly here, were considered by the DCIWG and ICCVAM during development of these recommended performance standards.

One commenter (Dr. Roland Roguet, L'ORÉAL Research) agreed with the importance of establishing test method performance standards and provided several suggestions for revising the performance standards for *in vitro* human skin model systems. In addition to minor editorial changes, the suggestions included noting that EPISKIN™ is currently commercially available for skin corrosivity assessments and that the test method can discriminate between United Nations (UN) transportation Packing Group I and Packing Group II/III substances. ICCVAM incorporated the suggested editorial changes and revised the text to note that EPISKIN™ is commercially available. However, the performance standards were not revised to reflect the ability of EPISKIN™ to discriminate between UN Packing Group I and Packing Group II/III substances. The ability of a test method to classify corrosive substances by UN Packing Group is applicable to the U.S. Department of Transportation (DOT) and not to other Federal agencies, and current U.S. DOT transportation regulations require that test method be able to classify the corrosive potential of substances across all three UN subcategories of corrosivity.

Two commenters (Drs. John Harbell and Rodger Curren, Institute for In Vitro Sciences, Inc.) commended the efforts of NICEATM and the DCIWG for drafting performance standards for the three types of *in vitro* skin corrosivity test methods and stated that these standards represent a substantial step forward in regulatory toxicology. Suggestions for improving the proposed performance standards included adding to the test report specifications the phrase "if relevant to the conduct of the study" where appropriate to make them less prescriptive and including specifications for test acceptance criteria (e.g., acceptable range of positive responses). The performance standards for all three test method systems were revised to include these suggested additions.

Another commenter (Mr. Troy Seidle, People for the Ethical Treatment of Animals [PETA]) stated that PETA appreciated the effort involved in the development of these performance standards and

that they were "hopeful that they would not only satisfy the needs of U.S. regulatory agencies, given their inability to lawfully require or recommend the use of proprietary test methods, but will also be useful in preventing future bottlenecks in the validation pipeline both domestically and internationally." The commenter indicated that PETA was in general agreement with the content of these documents. Additional comments were provided that did not pertain to the proposed performance standards but rather were directed at the manner in which *in vitro* test methods results are used in making decisions about the corrosivity/non-corrosivity of a test substance. Specifically, the commenter disagreed with ICCVAM's recommendation that *in vitro* human skin model systems (i.e., EpiDerm™ and EPISKIN™) (ICCVAM 2002) could be used as screening methods where substances inducing a positive result would be classified as corrosive, while substances inducing a negative result should undergo additional testing in accordance with the globally-harmonized tiered testing strategy (UN 2001). Rather, the commenter recommended that these *in vitro* tests should be considered as full replacements for the *in vivo* rabbit skin test method. ICCVAM's recommendation was based on the high false negative rates for these test methods (17% for EPISKIN™, 13% for EpiDerm™ [EPI-200], and 12% for TER) for identifying corrosive substances, and the irreversible permanent damage that could result from exposure to corrosive substances that were not properly classified and labelled as corrosive hazards. The commenter also noted that Worth et al. (1998) reported that the frequency of false negatives in the human skin model test systems, when combined with pH measurements and computerized structure-activity relationship (SAR) modeling in a sequential testing approach, could be reduced to zero. In response, ICCVAM noted that Worth et al. (1998) acknowledged that the SAR models used in this evaluation had not yet been validated. In addition, while eliminating test substances from further consideration based on pH alone appears useful, NICEATM could not locate a published standardized protocol for preparing solutions (in terms of the amount of the substance dissolved/mixed with water) for determining the pH. Furthermore, the approach also results in a relatively high percentage of false positives. Lacking a standardized protocol and validation of this approach, ICCVAM concluded that it is premature to formally evaluate this sequential testing strategy.

ICCVAM recognizes that, to date, a careful evaluation of the reliability and accuracy of the current *in vivo* rabbit skin corrosivity test has not yet been conducted. To correct this deficiency, NICEATM is reviewing *in vivo* rabbit skin corrosivity data that has been generated using current test method procedures. These data are being extracted from the published literature, and from data submitted by U.S. Federal regulatory agencies and commercial organizations to ICCVAM. Once these data have been tabulated and the appropriately analyzed, scientifically sound estimates of the reliability and under-prediction rates of the *in vivo* rabbit skin corrosivity test should be available which can be used to compare the performance of the *in vitro* corrosivity test methods.

A fourth commenter (Dr. Manfred Liebsch, Zentralstelle zur Erfassung und Bewertung von Ersatz- und Ergänzungsmethoden zum Tierversuch [German Centre for Documentation and Evaluation of Alternatives to Animal Experiments]; ZEBET) commented that ZEBET very much welcomed the general concept and the definition of performance standards for the future development of test systems that claim to be scientifically equivalent to existing validated systems. However, he recommended that ICCVAM adopt only the 12 chemicals provided in the proposed OECD test guidelines for the TER and *in vitro* human skin model systems as reference chemicals, rather

than the proposed 24 chemicals. He proposed that the other 12 reference chemicals could be recommended for test refinement if the first set of 12 reference chemicals were not classified 100% correctly.

ICCVAM considered these comments and decided to retain the current list of 24 chemicals based in part on the following reasons:

- In terms of the *in vitro* human skin model systems, ICCVAM agrees with the European Centre for the Validation of Alternative Methods (ECVAM) that 24 reference chemicals are needed, at a minimum, to adequately evaluate the reliability and accuracy of a test method that is mechanistically and functionally similar to EPISKIN™ (Liebsch et al. 2000). In addition, including substances that tested false positive or false negative in the validated reference test method in this minimum list allows the developer of a proposed test method to potentially demonstrate that their method provides a greater level of accuracy than the validated test method. However, ICCVAM concluded that 12 chemicals (6 noncorrosive and 6 corrosive) could be used by a naïve laboratory to demonstrate proficiency with the validated reference test methods (i.e. EPISKIN™, EpiDerm™ [EPI-200]) and should therefore be referred to as proficiency chemicals.
- In terms of the rat skin TER assay, the list of 24 reference chemicals was recommended for use only when a substantial protocol change was incorporated into the test method (e.g., skin from an animal of a different age, strain, and/or species than that the validated reference test method. In any case, 12 calibration chemicals would still be used to calibrate a rat skin TER assay in the hands of new investigator.
- ICCVAM did not include acrylic acid (Chemical Abstract Services Registry Number [CASRN] 79-10-7), one of the twelve reference chemicals recommend in the OECD test guidelines, in its list of reference chemicals for these two test methods because no data are available for this chemical in EPISKIN™ and the chemical was not used in the validation of EpiDerm™ (EPI-200).

The proposed ICCVAM performance standards for *in vitro* corrosivity test methods were discussed at the August 12-13, 2003 meeting of the SACATM (information about this meeting can be found at *http://iccvam.niehs.nih.gov/about/sacatm.htm*). The rationale and process for the development of the performance standards for these validated test methods was discussed, as was the recommended essential test method components, the minimum list of reference chemicals, and the specified levels of accuracy and reliability that should be achieved by a mechanistically and functionally similar test method. There was public comment on the proposed performance standards prior to the general discussion by the SACATM. The single public commenter at the meeting expressed opposition to the ICCVAM recommended application of these test methods as screening assays in a tiered testing strategy. NICEATM responded by reiterating that the recommended application was the consensus decision of an independent Peer Review Panel, the ICCVAM Corrosivity Working Group, and ICCVAM, and was based on the relatively high false negative rates of the *in vitro* test methods.

During the general discussion, the SACATM discussed the scientific validity of employing the proposed performance standards in the validation of new test methods that were mechanistically and functionally similar to the validated reference test method. The SACATM generally agreed that the approach should expedite the development and implementation of alternative test methods,

but that new (and unrelated) test methods would require more extensive validation, as outlined in the ICCVAM guidelines for test method validation. In addition, the SACATM emphasized that all test methods should be validated against the original animal-based reference test method and not the validated *in vitro* reference test method.

On October 28 and 29, 2003, an EPA Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) Scientific Advisory Panel (SAP) Meeting was held to review scientific issues being considered by the EPA regarding data quality for *in vitro* tests used as alternatives to animal studies for regulatory purposes (information about this meeting can be obtained at *http://www.epa.gov/scipoly/sap/*). During this meeting, the SAP reviewed the draft Performance Standards developed by ICCVAM for three types of *in vitro* corrosivity test systems. Based on its evaluation, the SAP:

- endorsed the Performance Standards approach to identify and validate *in vitro* test methods that are structurally and functionally similar to a validated *in vitro* reference test method
- concurred that the Performance Standards prepared by ICCVAM were very well described for each of the three tests
- concluded that the information generated using Performance Standards should provide a basis to determine whether a test is mechanistically and functionally similar to a validated *in vitro* test method

The SAP expressed the view that the strength of the Performance Standards approach to validating a new test method or those structurally and functionally similar to a validated reference test method derives from the stated selection criteria for the Reference Chemical set. The Panel also concluded that the approach of specifying a known level of accuracy and reliability for new test methods to meet in order to be considered equivalent to the validated reference test system was acceptable.

In terms of the Performance Standards, the EPA asked the SAP to comment on:

- the extent to which the essential test method components for each method adequately described the unique characteristics of the method necessary to determine whether a proposed test is mechanistically and functionally similar
- the strengths or weaknesses of this approach and any modifications to the criteria that should be considered
- whether test methods that are mechanistically and functionally similar to a validated reference test method should be demonstrated to be accurate for all of the chemicals in the Performance Standards

In responding to these questions, the SAP made a number of recommendations, several of which have been incorporated into the ICCVAM Performance Standards document. These include:

- provide examples of test methods that would be considered mechanistically and functionally similar to a validated reference test method
- expand the discussion on the use of benchmark controls
- specify the minimum replicate requirements for positive, negative, and benchmark controls
- provide general criteria for acceptance of concurrent positive controls in relation to historical positive control data

- expand the accuracy and reliability sections to include information on an effective approach for establishing intra- and inter-laboratory reproducibility
- clarify the concept of comparable performance characteristics for these three types of *in vitro* corrosivity screening assays

Several of the SAP recommendations were considered but could not be practically implemented within the scope of this document. One recommendation was that a single standard list of reference chemicals should be developed for validating all *in vitro* corrosivity test methods, while another was that a chemical repository should be established for reference samples/positive controls to be used by laboratories for developing/conducting *in vitro* skin studies. ICCVAM appreciates the value of a single standard list of reference chemicals for the validation of any test method proposed as an alternative to the traditional *in vivo* rabbit skin corrosivity test method. A similar approach was used by ICCVAM in its development of a proposed list of reference chemicals for the future validation of *in vitro* estrogen and androgen receptor binding and transcriptional activation assays (*http://iccvam.niehs.nih.gov/methods/endodocs/edfinrpt/edfinrpt.pdf*). Due to the sequence of events that led to the development of the Performance Standards for these three types of *in vitro* corrosivity test systems, this approach was not used. However, in the future, ICCVAM plans to routinely develop reference chemical sets for new test methods, which should eliminate the need for different reference chemical sets for test methods that measure or predict the same test endpoint. In response to the recommendation that ICCVAM establish a chemical repository, ICCVAM concluded that there was no need at the present time, considering that the recommended reference chemicals are commercially available.

The SAP recommended that laboratories should be allowed to determine their own positive control(s) and that the Performance Standards should not suggest specific examples. ICCVAM agreed, but considered that the inclusion of example positive controls for each test system would be useful to developers of new test methods. However, the appropriate sections in the Performance Standards document were revised to clarify that the indicated positive control chemical(s) is only provided as an example, .

The SAP raised concerns about the chemical classes represented in the list of reference chemicals, noting that some classes of potentially corrosive chemicals (e.g., hydrocarbons and halogenated hydrocarbons) were not represented. In addition, the Panel recommended including substances in the list of reference chemicals that would challenge the technical skill of laboratory staff. ICCVAM was unable to identify substances in the chemical classes of concern to the SAP that met the criteria established for inclusion as reference chemicals (see Section 1.2). Furthermore, although ICCVAM agrees that a wide range of chemicals, including those that are difficult to work with, would be ideal for determining limitations of a test method, the reference chemical list is limited to substances that meet the criteria described in Section 1.2. However, each reference chemical list includes substances that produced false negative and false positive responses in the validated reference test method (i.e., substances potentially difficult to work with).

The SAP questioned whether the numbers of chemicals in each list were sufficient for adequately demonstrating the reliability and accuracy of a proposed test method, even if it was mechanistically and functionally similar to a validated reference test method. ICCVAM agrees that this issue needs to be addressed further, especially for alternative test methods that are proposed as total

replacements for a traditional *in vivo* test method. However, since these *in vitro* dermal corrosivity test methods are recommended as screening tests, the minimum set of reference chemicals is considered appropriate for this purpose. Nevertheless, as more data regarding the limitations of these test methods for certain classes of chemicals or products are generated, or if improved methods are proposed as replacements, it may be desirable to revise the minimum reference list to ensure that these classes are better represented.

With regard to Quality Control issues, the SAP considered:

- the utility of and necessity for training or calibration sets in assuring data quality
- aspects of the quality control criteria that are necessary for assuring the integrity of such systems over time and from lot-to-lot
- the advantages and disadvantages of including concurrent positive and negative controls with *in vitro* assays when used as alternatives to animal testing
- whether benchmark controls serve a useful purpose to demonstrate the level of response that can be expected for each chemical class for each lot of proprietary test method assays

The SAP noted that individual test facilities may detect failures or out-of-specification performance of a proprietary test method and proceed according to their operating procedures, but the lack of a Good Manufacturing Process-like regulatory authority does not require these failures to be reported to and addressed by the vendor. Other facilities may unknowingly use an inadequate/underperforming proprietary test method without benefit of the experiences of the first facility. Thus, the SAP recommended that proprietary test method quality control reports should be compiled by the vendor and reported to purchasers of that test method. This issue was not specifically addressed in the Performance Standards, but will likely be addressed in an Organisation for Economic Co-operation and Development (OECD) Advisory Document on the Application of Good Laboratory Practices to *In Vitro* Studies that is currently under development.

The SAP commented on the essential role that concurrent positive control(s) have in ensuring the adequacy of *in vitro* studies, and on the appropriate role and properties that benchmark controls might have in such studies. The Performance Standards now provide greater emphasis on these topics.

Following issuance of the proposed performance standards for public comment, the DCIWG and ICCVAM revised some of the terminology in order to eliminate potential confusion. Specifically, “minimum performance standards” was revised to “performance standards” and “minimum procedural standards” was revised to “essential test method components”.

The following Sections describe the performance standards that should be met for three types of *in vitro* corrosivity test methods proposed for testing the skin corrosion hazard potential of chemicals (membrane barrier test systems, human skin model systems, and the rat skin TER test method). Validated versions of these three types of *in vitro* corrosivity test methods have been recommended by ICCVAM as screening assays for the detection of corrosive substances. ICCVAM recommends that proposed test methods that are mechanistically and functionally similar to the validated reference test methods must demonstrate comparable performance using the minimum list of reference substances included in these performance standards, and that decisions on comparable performance should be handled on a case-by-case basis. While it would be desirable for such

test methods to have reliability and accuracy values at least as good as that of the corresponding validated reference test method, some flexibility might be acceptable to the extent that it would not compromise the ultimate protection of human and animal health. For example, slightly higher false positive rates, while undesirable because they result in erroneous classification in a more hazardous category, do not result in lowered protection of human health. Because these test methods are used as *screening* tests, negative results will be followed with testing at least one animal as part of a dermal irritancy assessment, where false negatives should be detected by the presence of a corrosive skin lesion on the treated animal. Thus, a test method with a higher false negative rate will simply result in more positive corrosivity test results in the first animal tested for dermal irritancy. However, for future test methods proposed as *replacements* for existing test methods, minimum acceptable false positive and false negative rates will likely be recommended by ICCVAM, based on what is necessary to provide for equivalent or better protection of human or animal health or the environment.